Normal Forms for Description Logic Expressions of Clinical Concepts in SNOMED RT

Kent A. Spackman, M.D., Ph.D.
Oregon Health Sciences University, Portland, OR

Modern clinical terminologies organize concepts into multi-hierarchy structures that are defined by logic-based expressions, enabling compositional representation of clinical statements and supporting more complete and consistent retrieval of clinical data. The Systematized Nomenclature of Medicine, Reference Terminology (SNOMED© RT) gives each concept code a semantic definition stated in description logic. The process of development, testing and distribution of these definitions has highlighted the fact that a concept definition may take many different but logically equivalent forms, and has revealed a need for a set of ***normal forms*** *for authoring, distribution, and other purposes. This paper describes the difference between a choice of syntax and a choice of normal form, and defines several different normal forms, including a short canonical form, a long canonical form, and a distribution normal form.*

INTRODUCTION AND BACKGROUND

Structured clinical terminologies are designed to facilitate the storage and retrieval of clinical concepts from multiple perspectives rather than from a single perspective, such as that based on a code-oriented strict hierarchy. A new version of The Systematized Nomenclature of Medicine (SNOMED)[1], named SNOMED RT, has been created in order to improve its capabilities with respect to recent advances in clinical terminology practice, particularly the inclusion of multiple hierarchies and semantic definitions [2,3]. For several years there has been widespread acceptance of the idea that more formal representation of clinical meaning is necessary to fully exploit the capabilities of electronic records in improving and assuring the quality of health care [4,5,6,7,8]. In addition, it has been recognized that formal representation of the meaning of concepts in terminologies can lead to more consistency and quality within the terminologies themselves as well as more reliable and consistent retrieval of encoded clinical data [9,10].

CONCEPT DEFINITIONS

SNOMED RT has adopted a declarative semantics for concept definitions. *Declarative* semantics, as contrasted with *procedural* semantics, is the representation of meaning through expressions that can be understood without appeal to a special program or interpreter for manipulating those expressions. Declarative expressions most commonly use formal logic. Such methods have advanced significantly in the past decade, notably in the area known as description logic.

Description Logic

A description logic is a logic-based formalism for representing declarative semantics of terminological concepts[11,12]. Description logic statements denote the essential characteristics of concepts, that is, those characteristics that are always and necessarily true, and that serve to differentiate concepts from each other. Description logic statements represent essential characteristics using "concept-forming operators," typically including logical conjunction ($\sqcap$) and existentially quantified role restrictions ($\exists$R.C).

Supertypes, Attributes (Roles) and Values

In SNOMED RT, each concept definition consists of a single logical expression that is a conjunction of any number of concept identifiers, with or without a conjunction of a number of quantified attribute-value pairs. Each concept identifier in the definition designates a supertype of the concept being defined. Supertypes are more general concepts. For example, hernia repair is a supertype of inguinal hernia repair. Supertypes logically subsume their subtypes, and form a subsumption hierarchy. Supertypes are sometimes equivalently referred to as the "parent" concepts in an "is-a" hierarchy.

Attribute-value pairs, also known as role-value pairs, are also used to specify definitional aspects of a concept. They are preceded by a quantifier, which in the general case can be either the existential

quantifier (∃) or the universal quantifier (∀). In SNOMED RT, we have used only the existential quantifier. For example, a hernia repair can be defined by the attribute "direct morphology" with value "hernia." The values, such as "hernia", are concepts in their own right and exist in their own hierarchy with their own definitions.

Primitive vs Defined (Non-primitive)

Each concept definition is designated as either "primitive" or "defined". A primitive definition is one that lacks sufficient defining characteristics, in a logic sense, to algorithmically identify its subtypes. In other words, subtypes of a primitive concept must be explicitly stated to be subtypes. Non-primitive definitions, on the other hand, do have sufficient defining characteristics to logically identify all subtype concepts. For example, "bacterial infectious disease" can be fully defined as having supertype "infectious disease" and role-value pair "associated-etiology bacteria," since all bacterial infectious diseases will be subtypes of infectious disease and will have a value of "associated-etiology" that is bacteria or a subtype of bacteria.[1]

Syntax and Grammar for SNOMED Expressions

Before describing the proposed normal forms, we first briefly illustrate three options for syntax, in order to emphasize the difference between deciding on a syntax and deciding on a normal form. Description logic definitions in any normal form can be expressed in any valid syntax. Prior descriptions of SNOMED RT mentioned the KRSS syntax [2], leading to a misconception that SNOMED was based on KRSS. Instead, SNOMED RT description logic definitions were distributed in an XML syntax. In addition to providing yet another syntax, the XML document type definition (DTD) describes a grammar for SNOMED expressions.[2]

To illustrate their equivalence, a concept definition for "inguinal hernia repair" is given below in XML syntax, KRSS, and standard logic syntax.

XML syntax:

```
<cDef>
  <nm> Repair of inguinal hernia </nm>
  <cd>P1-B9810</cd>
  <defC>
   <cn>P1-00000</cn>
            -- surgical procedure
   <cn>P2-00030</cn>
            -- therapeutic procedure
  </defC>
  <defR>
    <rl>
      <some/><nm>SITE</nm>
      <val>T-D7040</val>
            -- inguinal canal
    </rl>
    <rl>
      <some/><nm>METHOD</nm>
      <val>P0-02087</val>
            -- surgical repair
    </rl>
    <rl>
      <some/><nm>DIRECT-MORPH</nm>
      <val>M-31500</val>
            -- hernia
    </rl>
  </defR>
</cDef>
```

KRSS syntax:

```
(defconcept P1-B9810
   (and P1-00000 P2-00030
      (some SITE T-D7040)
      (some METHOD P0-02087)
      (some DIRECT-MORPH M-31500)))
```

Standard logic syntax:

```
P1-B9810 ≐ P1-00000 ⊓ P2-00030 ⊓
           ∃SITE.T-D7040 ⊓
           ∃METHOD.P0-02087 ⊓
           ∃DIRECT-MORPH.M-31500
```

The symbol "≐" indicates logical equivalence, or in other words a fully defined, non-primitive concept definition. When specifying a primitive definition, the symbol "⊑" is used. For the remainder of this paper we will use the logic syntax because of its brevity.

Need for Normal Forms

Normal forms serve several purposes. First, they permit us to standardize the distribution format so that users know what content to expect when receiving a file containing concept definitions, regardless of the syntax. Second, naming the

[1] In order to deal with the special problems of representing anatomical partitive relationships, we used a transitive construct for part-whole relationships, virtually identical to the GALEN "specialised-by" construct. [13]

[2] The XML DTD, part of the SNOMED RT distribution, has a feature called "rolegroups" that allows roles to be grouped, but no definitions used this feature.

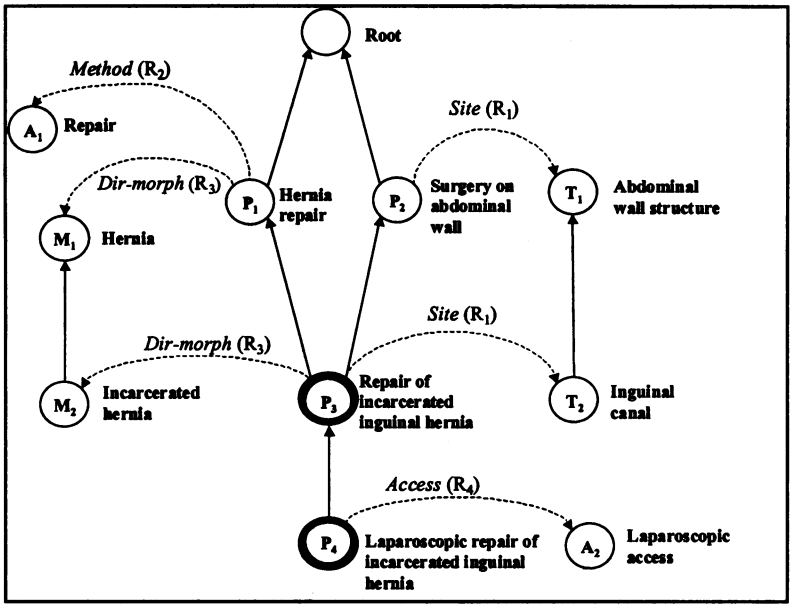

Figure 1: Graphical Representation of Definitions

different forms allows us to communicate to each other clearly and unambiguously, avoiding miscommunication and confusion. For example, during the concept definition process there has been confusion among modelers about which roles need to be explicitly modeled and which ones can be left unstated. Some of this confusion arises because of uncertainty about which roles and values are inherited from supertypes. The confusion also arises because of the multiple logically equivalent forms a definition can take, even though there is exactly one valid logical definition upon which the different forms are all based. Being able to name some of these different logically equivalent forms facilitates communication among modelers, reviewers, and implementers. Finally, having a canonical form can facilitate implementation by giving users a format that maximally decomposes expressions for recording, storing and retrieving both pre-coordinated concept descriptions and post-coordinated or compositional expressions.

A Relatively Simple Example

Figure 1 presents a graphical depiction of the concept "Laparoscopic repair of incarcerated inguinal hernia," labeled P_4, and its definitional relationships to other concepts in a mini-terminology. In this example, all concepts are primitive except P_3 and P_4, which are fully defined. Note that the structure of the diagram, and the analysis done here on canonical forms, assumes that all inferable logical relationships have been discovered algorithmically. In other words, the generation of canonical forms described here must be done after a description logic classifier has asserted all inferable subtype relationships.

Option 1: Include only the most proximate supertype(s)		
1A	$P_3 \sqcap \exists R_1.T_1 \sqcap \exists R_1.T_2 \sqcap \exists R_2.A_1 \sqcap \exists R_3.M_1 \sqcap \exists R_3.M_2 \sqcap \exists R_4.A_2$	
1B	$P_3 \sqcap \exists R_1.T_2 \sqcap \exists R_2.A_1 \sqcap \exists R_3.M_2 \sqcap \exists R_4.A_2$	**⇐ Distribution normal form**
1C	$P_3 \sqcap \exists R_4.A_2$	**⇐ Authoring normal form**
1D	$P_3 \sqcap \exists R_1.T_2 \sqcap \exists R_3.M_2 \sqcap \exists R_4.A_2$	
Option 2: Include only the most proximate *primitive* supertype(s)		
2A	$P_1 \sqcap P_2 \sqcap \exists R_1.T_1 \sqcap \exists R_1.T_2 \sqcap \exists R_2.A_1 \sqcap \exists R_3.M_1 \sqcap \exists R_3.M_2 \sqcap \exists R_4.A_2$	
2B	$P_1 \sqcap P_2 \sqcap \exists R_1.T_2 \sqcap \exists R_2.A_1 \sqcap \exists R_3.M_2 \sqcap \exists R_4.A_2$	**⇐ Long canonical form**
2C	$P_1 \sqcap P_2 \sqcap \exists R_4.A_2$	⇐ Not a correct definition of P_4
2D	$P_1 \sqcap P_2 \sqcap \exists R_1.T_2 \sqcap \exists R_3.M_2 \sqcap \exists R_4.A_2$	**⇐ Short canonical form**

Option A= Include all roles, redundant.
Option B= Include roles that differentiate from root, non-redundant.
Option C= Include roles that differentiate from the most proximate supertype(s)
Option D= Include roles that differentiate from the most proximate *primitive* supertype(s).

Table 1. Multiple logically equivalent forms for defining "Laparoscopic repair of incarcerated inguinal hernia". (Concept P_4 in Figure 1)

In Figure 1, "is-a" links from primitive concepts to the root have been omitted to keep the diagram simple.

Analyzing Options for Normal Forms

Several decisions can be made about what to include in a normal form. Since all definitions are merely conjunctions of supertypes and role-value pairs, these decisions can be broken down into decisions about which supertypes to include, and decisions about which roles to include.

Decisions about supertypes include whether to include redundant supertypes, and whether to include the most proximate supertypes or only the *most proximate primitive* supertypes. Decomposing non-primitive supertypes leads to determination of the most proximate primitive supertypes. Graphically this consists of moving upward in the hierarchy, along all paths, and stopping at primitive supertypes. Options for including supertypes in definitions are listed as options 1 and 2 in Table 1.

Decisions about roles include whether to include redundant roles, and whether to include roles that differentiate from the root, the most proximate supertype, or the most proximate primitive supertypes. These decisions are listed as options A, B, C and D in Table 1.

Combining the various options, Table 1 illustrates eight possible standard forms for the definition of P_4, seven of which are logically equivalent (2C is the exception -- it is logically not a correct definition of P_4).

Proposed Normal Forms

Based on the analysis illustrated in Table 1, we propose four different normal forms: authoring normal form, distribution normal form, short canonical form, and long canonical form.

Authoring normal form (option 1C in Table 1) is the most concise form and also has the advantage of being relatively more stable in the presence of changes to definitions higher in the hierarchy. Examining these various forms for the definition of P_4, "laparoscopic repair of inguinal hernia", one can see that the simplest expression is that in which one simply states that P_4 is a subtype of P_3 "repair of incarcerated inguinal hernia" with laparoscopic approach (option 1C in Table 1):

$$P_4 \doteq P_3 \sqcap \exists R_4.A_2$$

This is a concise definition, probably easiest to author and to obtain concurrence among independent reviewers of the definitions. However, another perspective would suggest that completely specifying the value for each role may avoid disagreement or confusion in the modeling process.

Distribution normal form explicitly states all the roles, including those that are inherited (option 1B in Table 1):

$$P_4 \doteq P_3 \sqcap \exists R_1.T_2 \sqcap \exists R_2.A_1 \sqcap \exists R_3.M_2 \sqcap \exists R_4.A_2$$

P_3 already logically implies the meaning of the first three roles; among roles, R_4 alone differentiates P_4 from P_3. The additional information, although logically unnecessary, avoids the need to examine supertypes to determine the role value for a concept. This form is the one used to populate our relational tables for distribution.

Both authoring form and distribution form explicitly list all (inferred) proximate supertypes, avoiding the necessity of recomputing the subtypes of fully defined concepts.

On the other hand, when comparing compositional expressions we may want to have a maximally decomposed form, also known as a *canonical form*. In canonical form, we maximally decompose a concept into its primitive defining supertypes. Because all logically equivalent expressions will decompose into the same set of primitives, canonical forms are very useful for comparing post-coordinated expressions. In order to do retrieval or analysis of stored clinical data, the stored expressions and the query expression would both be compared in their canonical form when evaluating equivalence or subsumption.

Short canonical form (option 2D in Table 1) is the least redundant decomposed (canonical) form, since it lists only the roles that differentiate the concept from its primitive supertypes. This is the form that was used for distributing XML-syntax definitions of concepts in SNOMED RT.

Long canonical form (option 2B in Table 1) has its supertypes decomposed, like short canonical form, but it lists values for all the roles, like distribution normal form. Definitions in this form are more "friendly" to tabular or relational representations of concepts. However, it should be noted that it is

entirely possible to have more than one role-value pair for the same role, neither of which subsumes the other. In other words, definitions in long canonical form cannot necessarily be put in a relational table with one column per role and one row per concept.

CONCLUSION

The proposed normal forms are intended to simplify communication about description logic definitions for clinical terminology. Combining these normal forms with the relatively simple XML grammar and syntax should make description-logic-based clinical terminology more accessible to both users and implementers of clinical systems.

Acknowledgments

This work is supported in part by a grant from the College of American Pathologists. Thanks go to Bob Dolin for his thoughtful comments on a draft of the paper; to Eric Mays for his encouragement that a description of normal forms would be of interest; and to the SNOMED modelers and beta testers for expressing the need for more clarity about definitional forms, syntax, and grammar.

References

1. Cote RA, Rothwell DJ, Palotay JL, Beckett RS, Brochu L, eds. The Systematized Nomenclature of Human and Veterinary Medicine: SNOMED International. Northfield, IL: College of American Pathologists, 1993.

2. Spackman KA, Campbell KE, Cote RA. SNOMED RT: A reference terminology for health care. In: AMIA Annual Fall Symposium, 640-644, 1997.

3. Spackman KA, et al, eds. SNOMED® RT: Systematized Nomenclature of Medicine, Reference Terminology. Northfield, IL: College of American Pathologists, 2000.

4. Campbell KE, Das AK, Musen MA. A Logical foundation for representation of clinical data. J Am Med Informatics Assoc 1:218-232, 1994.

5. Evans DA, Cimino JJ, Hersh WR, Huff SM, Bell DS. Toward a medical concept representation language. J Am Med Inform Assoc. 1994; 1: 207-17.

6. Rector AL, Bechhofer S, Goble CA, et al. The GRAIL concept modelling language for medical terminology. Artificial Intelligence in Medicine 9(2):139-171, 1997

7. Rector AL, Nowlan WA, Kay S. Conceptual knowledge: the core of medical information systems. In Lun KC et al (Eds): Proceedings of the Seventh World Congress on Medical Informatics (MEDINFO '92). Amsterdam, North Holland, 1992: 1420-6.

8. O'Neil MJ, Payne C, Read JD. Read Codes Version 3: A User Led Terminology. Meth Inform Med 1995; 34: 187-92

9. Schulz EB, Barrett JW, Price C. Read Code Quality Assurance: From Simple Syntax to Semantic Stability. J Am Med Inform Assoc. 1998; 5: 337-46.

10. Schulz EB, Barrett JW, Price C. Semantic Quality through Semantic Definition: Refining the Read Codes through Internal Consistency. In Masys DR (Ed). Proceedings of the 1997 AMIA Annual Fall Symposium. Philadelphia: Hanley & Belfus, 1997: 615-619.

11. Woods WA, Schmolze JG. The KL-ONE family. Computers and Mathematics with Applications -- Special Issue on Artificial Intelligence, 23(2-5):133-177, 1992.

12. Mays E, Dionne R, Weida R. K-REP system overview. SIGART Bulletin 2(3):88-92, 1991.

13. Rector AL, Bechhofer S, Goble C, Horrocks I, Nowlan A, Solomon WD. The GRAIL concept modelling language for medical terminology. Artificial Intelligence in Medicine 9(2):139, 1997